Independencies in Bayesian Networks

Independence

- For events $a, b, a \perp b \in P$ if :
  - $P(a, b) = P(a)P(b)$
  - $P(a|b) = P(a)$
  - $P(b|a) = P(b)$
- For random variables $X, Y, X \perp Y \in P$ if:
  - $P(X, Y) = P(X)P(Y)$
  - $P(X|Y) = P(X)$
  - $P(Y|X) = P(X)$

Conditional Independence

- for RVs X, Y, Z, $(X \perp Y|Z)$ if:
  - $P(X, Y|Z) = P(X|Z)P(Y|Z)$
  - $P(X|Z, Y) = P(X|Z)$
  - $P(Y|Z, X) = P(Y|Z)$
  - $P(X, Y, Z) \propto \phi_1(X, Z)\phi_2(Y, Z)$

Independence & Factorization

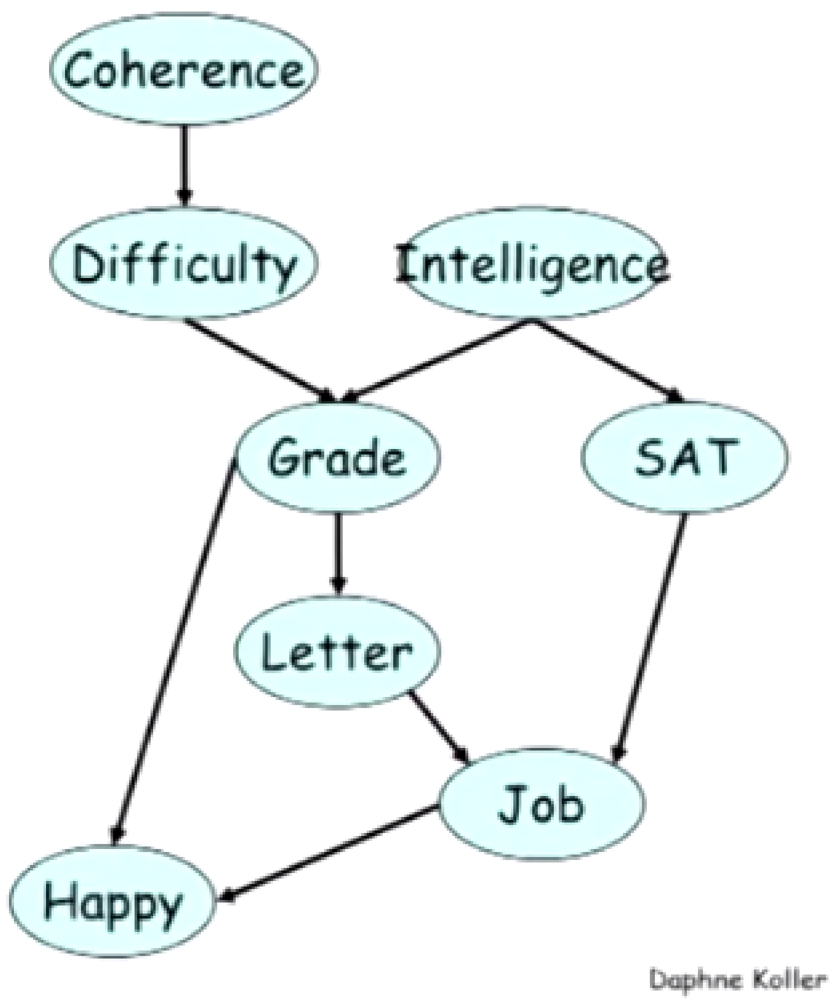
Factorization of a distribution P implies independency that hold in P. Then, if P factorizes over G, can we read these independency from the structure of G?

Flow of Influence & d-separation

Definition: X and Y are d-separated in G given Z if there is no active trail in G between X and Y given Z. Notation: $d - sep_G(X, Y|Z)$

Theorem: If P factorizes over G, and $d - sep_G(X, Y|Z)$ then P satisfies $(X \perp Y|Z)$

Any node is d-separated from its non-descendants given its parents. For example, in following graph, the RV Letter is only determined by its parent Grade. If we give Grade and SAT at same time, there is no active trail from SAT to Letter.



From above sentence, we can conclude that if P factorizes over G, then in P, any RV is independent of its non-descendants given its parents.

I-Maps

From d-separation in G, we can have P satisfies corresponding independence statements. Then we define $I(G)$ as the d-separation in G. $I(G) = \{(X \perp Y|Z) : d - sep_G(X, Y|Z)\}$

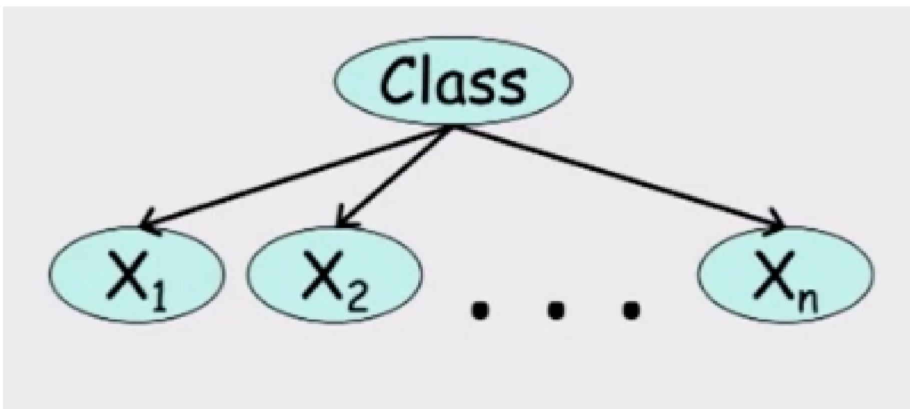
Definition: If P satisfies $I(G)$ we say that G is an I-map of P. In other word, the independence relationship in G is subset of independence relationship in P.

Factorization to Independence in BNs

Theorem: If P factorizes over G, then G is an I-map for P.

Theorem: If G is an I-map for P, then P factorizes over G. This factorization may not be optimal, since some independences might not be captured by G.

Naive Bayes



From the graph we can see that the Naive Bayes Model make assumption that given class C, all features $X_1, ...X_n$ are independent. Which means all features are conditional independent.

From the assumption, we have $P(C, X_1, ..., X_n) = P(C) \prod_{i=1}^{n} P(X_i|C)$

Summary: - Simple approach for classification, easy computation and easy construct - Effective in domains with many weakly relevant features - Reduce performance when many features are strongly correlated